Statistics and Probability Letters 150 (2019) 88–93

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The variance of the average depth of a pure birth process converges to 7

Ken R. Duffy [a], Gianfelice Meli [a,*], Seva Shneer [b]

[a] Hamilton Institute, Maynooth University, Maynooth, Ireland
[b] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK

ARTICLE INFO

Article history:
Received 21 December 2018
Received in revised form 26 February 2019
Accepted 26 February 2019
Available online 5 March 2019

Keywords:
Pure birth process
Variance of the average depth

MSC:
92D25
60J85

ABSTRACT

If trees are constructed from a pure birth process and one defines the depth of a leaf to be the number of edges to its root, it is known that the variance in the depth of a randomly selected leaf of a randomly selected tree grows linearly in time. In this letter, we instead consider the variance of the average depth of leaves within each individual tree, establishing that, in contrast, it converges to a constant, 7. This result indicates that while the variance in leaf depths amongst the ensemble of pure birth processes undergoes large fluctuations, the average depth across individual trees is much more consistent.

1. Introduction

Continuous time branching processes form fundamental building blocks of many stochastic models (e.g. Kimmel and Axelrod (2002)) and much is known about many statistics associated with them. A pure birth process (Resnick, 2013) is the simplest continuous time branching process. It describes the growth of a directed tree that starts at time 0 with a root, which is the first leaf. Each leaf extends the tree by creating two new leaves after an exponentially distributed time with mean $1/\lambda$, independently of everything else. Pure birth processes appear as a fundamental model of study in a large number of applications from data structures in computer science to likelihood methods in phylogenetics to the study of random walkers on random graphs, and are well studied.

Of interest to us here is a measure of tree depth, the distance from root to leaves. If one conditions on the number of nodes, much is known. For example, Pittel (1984) linked prior results regarding binary search trees (Robson, 1979; Flajolet and Odlyzko, 1980; Devroye, 1986) to continuous time Markovian branching processes, establishing scaling properties of the depth of both the shortest and longest leaf. Further extensions of those results have since been found (Pittel, 1994; Biggins and Grey, 1997). Without conditioning on the number of nodes in the tree, relatively little appears in the literature. For a pure birth process, it is known that the mean depth of a randomly chosen leaf in a randomly selected tree grows as $2\lambda t$ with variance $2\lambda t$ (Samuels, 1971). However, for many applications, particularly in the life sciences e.g. Perié et al. (2014) and Marchingo et al. (2016), one is interested in the properties of individual growing trees. Denoting the number of leaves in a random tree at time t by $Z(t)$ and the sum of their depths by $G(t)$, with $Z(0) = 1$ and $G(0) = 0$. The object of the present study is the variance across trees of the average depth of the leaves within them, i.e. $G(t)/Z(t)$, and our main result is as follows.

* Corresponding author.
E-mail address: gianfelice.meli@mu.ie (G. Meli).

https://doi.org/10.1016/j.spl.2019.02.015
0167-7152/

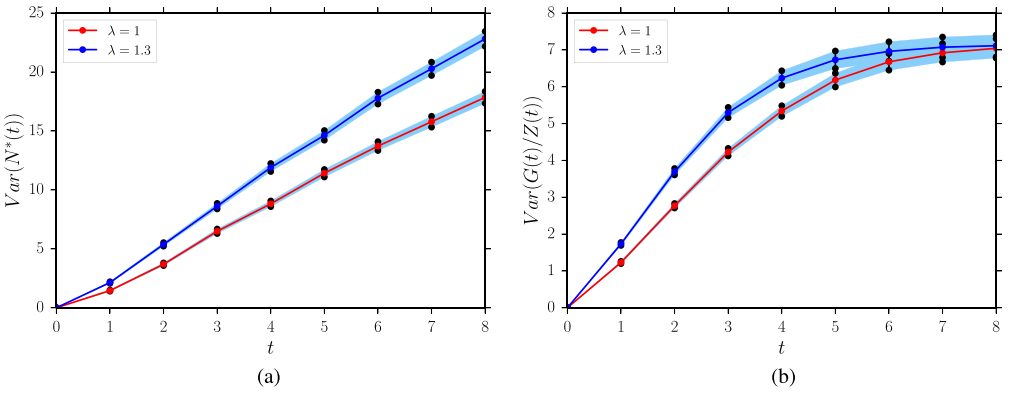

Fig. 1. 10^4 Monte Carlo simulations of a pure birth process were used to determine the variance of the average depth of a leaf in a random tree, $N^*(t)$, and the variance of the average depth of a tree, $G(t)/Z(t)$, where λ equals 1 (lower lines) and 1.3 (upper lines), and the blue shaded region indicates 95% confidence intervals based on bootstrap percentiles (Efron and Tibshirani, 1994, Chapter 13). (a) Consistent with Samuels (1971), $\mathrm{Var}\,(N^*(t)) \sim 2\lambda t$. (b) Consistent with Theorem 1, $\mathrm{Var}\,(G(t)/Z(t)) \sim 7$ irrespective of λ.

Theorem 1. *For a pure birth process, we have that*

$$\lim_{t\to\infty} Var\left(\frac{G(t)}{Z(t)}\right) = 7.$$

In addition to the results in Samuels (1971), this finding is potentially surprising because it is known that the two processes $\{Z(t)\}$ and $\{G(t)\}$ have different growth rates, $e^{\lambda t}$ and $te^{\lambda t}$, respectively (Jagers, 1969; Weber et al., 2016), from which one might anticipate that the variability of the average depth of a tree diverges to infinity as t^2. Those suppositions are incorrect as it has recently been established that, for general continuous time branching processes, $Z(t)$ and $G(t)$ are strongly correlated at the level of sample paths (Meli et al., 2018), and that, for a pure birth process, $\lim_t Z(t)/(tG(t)) = 2\lambda$ almost surely. A visualization of the result in Theorem 1, obtained by Monte Carlo simulation, is provided in Fig. 1. Note that the result does not depend on λ, which only influences the speed of convergence.

In order to evaluate $\mathrm{Var}\,(G(t)/Z(t))$, we condition the average generation $G(t)/Z(t)$ on the number of leaves at time t, $Z(t)$. By the Law of Total Variance (e.g. Blitzstein and Hwang (2014))

$$\mathrm{Var}\left(\frac{G(t)}{Z(t)}\right) = \mathbb{E}\left(\mathrm{Var}\left(\frac{G(t)}{Z(t)}\middle|Z(t)\right)\right) + \mathrm{Var}\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}\middle|Z(t)\right)\right) \tag{1}$$

and, in order to study the variance of the average depth of the leaves at time t, we study the quantities $\mathbb{E}\,(G(t)/Z(t)|Z(t))$ and $\mathrm{Var}\,(G(t)/Z(t)|Z(t))$ in Lemmas 2 and 3, respectively. Theorem 1 then follows.

2. Results

Before proceeding with the analysis of the two terms on the RHS of (1), we prove a lemma that will simplify the proofs of Lemmas 2 and 3. For that, we introduce a new process, $\{S(t)\}$, denoting the sum of the squares of the depths of the leaves at time t, which appears when the second moment of $G(t)/Z(t)$ is studied. In the following we also consider the discrete-time process associated with $\{G(t)\}$ and $\{S(t)\}$, namely $\{G_k\}$ and $\{S_k\}$, which account for the sum and the sum of the squares of the depths of the leaves, respectively, when the number of leaves is k.

Lemma 1. *We have that*

$$\mathbb{E}\left(\frac{G(t)}{Z(t)}\middle|Z(t)=k\right) = \frac{\mathbb{E}\,(G_k)}{k} = 2\sum_{i=2}^{k}\frac{1}{i}, \qquad \frac{\mathbb{E}\,(S_k)}{k} = 4\sum_{i=2}^{k-1}\frac{\mathbb{E}\,(G_i)}{i(i+1)} + \frac{\mathbb{E}\,(G_k)}{k}, \tag{2}$$

$$\mathbb{E}\left(\frac{G(t)^2}{Z(t)^2}\middle|Z(t)=k\right) = \frac{\mathbb{E}\left(G_k^2\right)}{k^2} = \frac{k+1}{k}\left(\sum_{i=1}^{k-1}\frac{\mathbb{E}\,(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{1}{(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{\mathbb{E}\,(G_i)}{i(i+2)}\right). \tag{3}$$

Proof. Throughout this proof, we condition on $Z(t) = k$ and denote by $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$ the depth of the k leaves present at time t, which are not independent. From the definitions, we have $G_k := \sum_{i=1}^{k}\Gamma_i$ and $S_k := \sum_{i=1}^{k}\Gamma_i^2$. The idea of the proof is to recover the formulas given above by finding recurrence equations for $\mathbb{E}\,(G_k)$, $\mathbb{E}\,(S_k)$, and $\mathbb{E}\left(G_k^2\right)$.

For $j \in \{1, 2, \ldots, k\}$, denote by I_j a random variable that takes value 1 if the jth leaf is the first one, among the k existing, to extend the tree with two new leaves, and 0 otherwise. The random variables in the set $\{I_j, \Gamma_1, \ldots, \Gamma_k\}$ are

independent for $j \in \{1, 2, \ldots, k\}$ and, due to the memoryless property of the exponential distribution, $\mathbb{P}\left(I_j = 1\right) = 1/k$ for all $j \in \{1, 2, \ldots, k\}$, with k the number of leaves in the tree. Furthermore, the I_j are not independent of each other because only one of them can assume value 1, i.e. $\sum_{j=1}^{k} I_j = 1$, implying that $I_j^2 = I_j$ and $I_j I_\ell = 0$ if $j \neq \ell$. With that in mind, we establish the following relations

$$G_{k+1} = G_k + \sum_{j=1}^{k} I_j \Gamma_j + 2, \qquad S_{k+1} = S_k + \sum_{j=1}^{k} I_j \left(2(\Gamma_j + 1)^2 - \Gamma_j^2\right), \tag{4}$$

$$G_{k+1}^2 = G_k^2 + \left(\sum_{j=1}^{k} I_j \Gamma_j\right)^2 + 4 + 4G_k + 4\sum_{j=1}^{k} I_j \Gamma_j + 2G_k \sum_{j=1}^{k} I_j \Gamma_j = G_k^2 + \sum_{j=1}^{k} I_j \Gamma_j^2 + 4 + 4G_k + 4\sum_{j=1}^{k} I_j \Gamma_j + 2G_k \sum_{j=1}^{k} I_j \Gamma_j. \tag{5}$$

From the first equation in (4) we obtain

$$\begin{aligned}\mathbb{E}\left(G_{k+1}\right) &= \mathbb{E}\left(G_k\right) + \sum_{j=1}^{k} \mathbb{E}\left(I_j \Gamma_j\right) + 2 = \mathbb{E}\left(G_k\right) + \sum_{j=1}^{k} \mathbb{E}\left(I_j\right) \mathbb{E}\left(\Gamma_j\right) + 2 \\ &= \mathbb{E}\left(G_k\right) + \frac{1}{k}\mathbb{E}\left(\sum_{j=1}^{k} \Gamma_j\right) + 2 = \mathbb{E}\left(G_k\right) + \frac{1}{k}\mathbb{E}\left(G_k\right) + 2 = \frac{k+1}{k}\mathbb{E}\left(G_k\right) + 2,\end{aligned}$$

where we have used that I_j and Γ_j are independent. This gives the following recurrence relation $\mathbb{E}\left(G_{k+1}\right)(k+1)^{-1} = \mathbb{E}\left(G_k\right)k^{-1} + 2(k+1)^{-1}$, that, solved with initial condition $\mathbb{E}\left(G_1\right) = 0$, results in the first formula in (2).

Similarly, using the second equation in (4), we have that

$$\begin{aligned}\mathbb{E}\left(S_{k+1}\right) &= \mathbb{E}\left(S_k\right) + \sum_{j=1}^{k} \mathbb{E}\left(I_j\right) \mathbb{E}\left(2(\Gamma_j + 1)^2 - \Gamma_j^2\right) = \mathbb{E}\left(S_k\right) + \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\left(\Gamma_j^2 + 4\Gamma_j + 2\right) \\ &= \mathbb{E}\left(S_k\right) + \frac{1}{k}\mathbb{E}\left(S_k\right) + \frac{4}{k}\mathbb{E}\left(G_k\right) + 2 = \frac{k+1}{k}\mathbb{E}\left(S_k\right) + \frac{4}{k}\mathbb{E}\left(G_k\right) + 2,\end{aligned}$$

from which we get the recurrence equation $\mathbb{E}\left(S_{k+1}\right)(k+1)^{-1} = \mathbb{E}\left(S_k\right)k^{-1} + 4\mathbb{E}\left(G_k\right)\left(k(k+1)\right)^{-1} + 2(k+1)^{-1}$. Solving this recursion with $\mathbb{E}\left(S_1\right) = \mathbb{E}\left(G_1\right) = 0$, we obtain the second result in (2).

Using (5) and the two results just found (i.e. the formulas in (2)), we can now find an expression for $\mathbb{E}\left(G_k^2\right)$.

$$\begin{aligned}\mathbb{E}\left(G_{k+1}^2\right) =& \mathbb{E}\left(G_k^2\right) + \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}\left(\Gamma_j^2\right) + 4 + 4\mathbb{E}\left(G_k\right) + \frac{4}{k}\sum_{j=1}^{k} \mathbb{E}\left(\Gamma_j\right) + 2\mathbb{E}\left(G_k\left(\sum_{j=1}^{k} I_j \Gamma_j\right)\right) \\ =& \mathbb{E}\left(G_k^2\right) + \frac{1}{k}\mathbb{E}\left(S_k\right) + 4 + \left(4 + \frac{4}{k}\right)\mathbb{E}\left(G_k\right) + \frac{2}{k}\mathbb{E}\left(G_k \sum_{j=1}^{k} \Gamma_j\right) \\ =& \mathbb{E}\left(G_k^2\right) + \frac{\mathbb{E}\left(S_k\right)}{k} + 4 + \frac{4(k+1)}{k}\mathbb{E}\left(G_k\right) + \frac{2}{k}\mathbb{E}\left(G_k^2\right) = \frac{k+2}{k}\mathbb{E}\left(G_k^2\right) + \frac{\mathbb{E}\left(S_k\right)}{k} + 4 + \frac{4(k+1)}{k}\mathbb{E}\left(G_k\right).\end{aligned}$$

The equation above can be rewritten as the recurrence equation

$$\frac{\mathbb{E}\left(G_{k+1}^2\right)}{(k+1)(k+2)} = \frac{\mathbb{E}\left(G_k^2\right)}{k(k+1)} + \frac{\mathbb{E}\left(S_k\right)}{k(k+1)(k+2)} + \frac{4}{(k+1)(k+2)} + \frac{4\mathbb{E}\left(G_k\right)}{k(k+2)},$$

that, when solved with initial condition $\mathbb{E}\left(G_1\right) = \mathbb{E}\left(G_1^2\right) = \mathbb{E}\left(S_1\right) = 0$, gives (3). □

We now use Lemma 1 to study the limit behaviour of the first term on the RHS of (1).

Lemma 2. *For a pure birth process, we have that*

$$\lim_{t\to\infty} \mathbb{E}\left(Var\left(\frac{G(t)}{Z(t)}\middle| Z(t)\right)\right) = 7 - \frac{2}{3}\pi^2.$$

Proof. Given that $\lim_{t\to\infty} Z(t) = \infty$ a.s. Harris (1963, Chapter 5), for every fixed $k \in \mathbb{N}$ we have that $\lim_{t\to\infty} \mathbb{P}(Z(t) = k) = 0$. This implies that

$$\lim_{t\to\infty} \mathbb{E}\left(\mathrm{Var}\left(\frac{G(t)}{Z(t)}\Big|Z(t)\right)\right) = \lim_{t\to\infty}\sum_{k=1}^{\infty}\mathrm{Var}\left(\frac{G_k}{k}\Big|Z(t)=k\right)\mathbb{P}(Z(t)=k) = \lim_{k\to\infty}\mathrm{Var}\left(\frac{G_k}{k}\right).$$

Using Lemma 1, we can now compute this variance:

$$\begin{aligned}
\mathrm{Var}\left(\frac{G_k}{k}\right) &= \frac{\mathbb{E}(G_k^2)}{k^2} - \frac{\mathbb{E}(G_k)^2}{k^2} = \frac{k+1}{k}\left[\sum_{i=1}^{k-1}\frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{1}{(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{\mathbb{E}(G_i)}{i(i+2)}\right] - \frac{\mathbb{E}(G_k)^2}{k^2}\\
&= \frac{k+1}{k}\left[\sum_{i=1}^{k-1}\frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{1}{(i+1)(i+2)} + 4\sum_{i=1}^{k}\frac{\mathbb{E}(G_i)}{i^2} + 4\sum_{i=1}^{k-1}\left(\frac{\mathbb{E}(G_i)}{i(i+2)} - \frac{\mathbb{E}(G_i)}{i^2}\right) - 4\frac{\mathbb{E}(G_k)}{k^2}\right]\\
&\quad - \frac{\mathbb{E}(G_k)^2}{k^2}\\
&= \frac{k+1}{k}\left[\sum_{i=1}^{k-1}\frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{1}{(i+1)(i+2)} + 4\sum_{i=2}^{k}\frac{1}{i^2} + \frac{\mathbb{E}(G_k)^2}{k^2} - 8\sum_{i=1}^{k-1}\frac{\mathbb{E}(G_i)}{i^2(i+2)} - 4\frac{\mathbb{E}(G_k)}{k^2}\right]\\
&\quad - \frac{\mathbb{E}(G_k)^2}{k^2}\\
&= \frac{k+1}{k}\left[\sum_{i=1}^{k-1}\frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{k-1}\frac{1}{(i+1)(i+2)} + 4\sum_{i=2}^{k}\frac{1}{i^2} - 8\sum_{i=1}^{k-1}\frac{\mathbb{E}(G_i)}{i^2(i+2)} - 4\frac{\mathbb{E}(G_k)}{k^2}\right] + \frac{\mathbb{E}(G_k)^2}{k^3},
\end{aligned}$$

where in the third equality we have added and subtracted the quantity

$$4\sum_{i=1}^{k}\frac{\mathbb{E}(G_i)}{i^2} = 8\sum_{i=2}^{k}\frac{1}{i}\sum_{j=2}^{i}\frac{1}{j} = 8\sum_{i=2}^{k}\frac{1}{i^2} + 8\sum_{i=2}^{k}\sum_{j=2}^{i-1}\frac{1}{ij} = 8\sum_{i=2}^{k}\frac{1}{i^2} + 4\left(\left(\sum_{i=2}^{k}\frac{1}{i}\right)^2 - \sum_{i=2}^{k}\frac{1}{i^2}\right) = 4\sum_{i=2}^{k}\frac{1}{i^2} + \frac{\mathbb{E}(G_k)^2}{k^2}.$$

Taking the limit as $k \to \infty$, we have that

$$\begin{aligned}
\lim_{k\to\infty}\mathrm{Var}\left(\frac{G_k}{k}\right) &= \sum_{i=1}^{\infty}\frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 4\sum_{i=1}^{\infty}\frac{1}{(i+1)(i+2)} + 4\sum_{i=2}^{\infty}\frac{1}{i^2} - 8\sum_{i=1}^{\infty}\frac{\mathbb{E}(G_i)}{i^2(i+2)}\\
&= \sum_{i=1}^{\infty}\frac{\mathbb{E}(S_i)}{i(i+1)(i+2)} + 2 + 4\left(\frac{\pi^2}{6} - 1\right) - 8\sum_{i=1}^{\infty}\frac{\mathbb{E}(G_i)}{i^2(i+2)}. \qquad (6)
\end{aligned}$$

Using Lemma 1, the first term on the RHS of (6) becomes

$$\begin{aligned}
&\sum_{i=1}^{\infty}\frac{1}{(i+1)(i+2)}\left(4\sum_{k=2}^{i-1}\frac{1}{k+1}\frac{\mathbb{E}(G_k)}{k} + \frac{\mathbb{E}(G_i)}{i}\right) = 8\sum_{i=1}^{\infty}\frac{1}{(i+1)(i+2)}\sum_{k=2}^{i-1}\frac{1}{k+1}\sum_{j=2}^{k}\frac{1}{j}\\
&\quad + 2\sum_{i=1}^{\infty}\frac{1}{(i+1)(i+2)}\sum_{k=2}^{i}\frac{1}{k}. \qquad (7)
\end{aligned}$$

The first term on the RHS of (7) is given by

$$8\sum_{j=2}^{\infty}\frac{1}{j}\sum_{k=j}^{\infty}\frac{1}{k+1}\sum_{i=k+1}^{\infty}\frac{1}{(i+1)(i+2)} = 8\sum_{j=2}^{\infty}\frac{1}{j}\sum_{k=j}^{\infty}\frac{1}{k+1}\frac{1}{k+2} = 8\sum_{j=2}^{\infty}\frac{1}{j}\frac{1}{j+1} = 4,$$

whereas the second one is given by

$$2\sum_{k=2}^{\infty}\frac{1}{k}\sum_{i=k}^{\infty}\frac{1}{(i+1)(i+2)} = 2\sum_{k=2}^{\infty}\frac{1}{k}\frac{1}{k+1} = 1.$$

So, the first sum in the RHS of (6) is equal to $4 + 1 = 5$. For the last sum in the RHS of (6), we have

$$-8\sum_{i=1}^{\infty}\frac{1}{i(i+2)}\frac{\mathbb{E}(G_i)}{i} = -16\sum_{i=1}^{\infty}\frac{1}{i(i+2)}\sum_{j=2}^{i}\frac{1}{j} = -16\sum_{j=2}^{\infty}\frac{1}{j}\sum_{i=j}^{\infty}\frac{1}{i(i+2)} = -16\sum_{j=2}^{\infty}\frac{1}{j}\frac{1+2j}{2j(j+1)} = -\frac{4}{3}(\pi^2 - 3).$$

Joining all these results, we obtain

$$\lim_{k\to\infty} \mathrm{Var}\left(\frac{G_k}{k}\right) = 5 + 2 + 4\left(\frac{\pi^2}{6} - 1\right) - \frac{4}{3}(\pi^2 - 3) = 7 - \frac{2}{3}\pi^2. \quad \square$$

Lemma 1 allows us to also understand the behaviour of the conditional variance of the expected average depth of the leaves given their number.

Lemma 3. *For a pure birth process, we have that*

$$\lim_{t\to\infty} Var\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}|Z(t)\right)\right) = \frac{2}{3}\pi^2 \approx 6.58.$$

Proof. From Lemma 1 we know that

$$\mathrm{Var}\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}|Z(t)\right)\right) = \mathrm{Var}\left(2\sum_{i=2}^{Z(t)}\frac{1}{i}\right) = 4\mathrm{Var}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right) = 4\left(\mathbb{E}\left(\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)^2\right) - \left(\mathbb{E}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)\right)^2\right), \tag{8}$$

where, in the second inequality, we have used the fact that the variance of a process does not change when a constant is added. Given that $\{Z(t)\}$ is a pure birth process, the distribution of $Z(t)$ is given by (e.g. Resnick (2013, pg. 430))

$$\mathbb{P}(Z(t) = k) = e^{-\lambda t}(1 - e^{-\lambda t})^{k-1}, \quad k = 1, 2, \ldots$$

where $1/\lambda$ is the expected time before a leaf generates two new leaves, which allows us to evaluate the second term in (8) exactly:

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right) &= \sum_{k=1}^{\infty}\mathbb{P}(Z(t)=k)\sum_{i=1}^{k}\frac{1}{i} = \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{k=1}^{\infty}(1-e^{-\lambda t})^k\sum_{i=1}^{k}\frac{1}{i}\\ &= \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{i=1}^{\infty}\frac{1}{i}\sum_{k=i}^{\infty}(1-e^{-\lambda t})^k = \frac{1}{(1-e^{-\lambda t})}\sum_{i=1}^{\infty}\frac{1}{i}(1-e^{-\lambda t})^i.\end{aligned}$$

Let $f(t) := \sum_{i=1}^{\infty}(1-e^{-\lambda t})^i/i$. Then

$$f'(t) = \lambda e^{-\lambda t}\sum_{i=1}^{\infty}\frac{1}{i}i(1-e^{-\lambda t})^{i-1} = \lambda e^{-\lambda t}\sum_{i=1}^{\infty}(1-e^{-\lambda t})^{i-1} = \lambda,$$

and, given $f(0) = 0$, we have that $f(t) = \lambda t$. This implies that

$$\mathbb{E}\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right) = \frac{\lambda t}{(1-e^{-\lambda t})} = \lambda t + o(1), \tag{9}$$

and the second term in the brackets on the RHS of (8) is therefore $(\lambda^2 t^2)/(1-e^{-\lambda t})^2$.

Consider the first term on the RHS of (8).

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{i=1}^{Z(t)}\frac{1}{i}\right)^2\right) &= \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{i=1}^{\infty}\left(\sum_{k=1}^{i}\frac{1}{k}\right)^2(1-e^{-\lambda t})^i\\ &= \frac{e^{-\lambda t}}{1-e^{-\lambda t}}\left(\sum_{i=1}^{\infty}\sum_{k=1}^{i}\frac{1}{k^2}(1-e^{-\lambda t})^i + 2\sum_{i=1}^{\infty}\sum_{k=1}^{i}\sum_{j=1}^{k-1}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^i\right).\end{aligned} \tag{10}$$

The first term in the brackets on the RHS of (10) is given by

$$\sum_{i=1}^{\infty}\sum_{k=1}^{i}\frac{1}{k^2}(1-e^{-\lambda t})^i = \sum_{k=1}^{\infty}\sum_{i=k}^{\infty}\frac{1}{k^2}(1-e^{-\lambda t})^i = e^{\lambda t}\sum_{k=1}^{\infty}\frac{1}{k^2}(1-e^{-\lambda t})^k.$$

For the second term, we have that

$$2\sum_{i=1}^{\infty}\sum_{k=1}^{i}\sum_{j=1}^{k-1}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^i = 2\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}\sum_{k=i}^{\infty}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^i = 2e^{\lambda t}\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}\frac{1}{k}\frac{1}{j}(1-e^{-\lambda t})^k.$$

Denoting with $g(t) := 2\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}(1-e^{-\lambda t})^k/(kj)$ and noticing that $g(0) = 0$ and

$$g'(t) = \frac{2\lambda e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{k=1}^{\infty}\sum_{j=1}^{k-1}\frac{1}{j}(1-e^{-\lambda t})^k = \frac{2\lambda e^{-\lambda t}}{1-e^{-\lambda t}}\sum_{j=1}^{\infty}\sum_{k=j+1}^{\infty}\frac{1}{j}(1-e^{-\lambda t})^k$$
$$= \frac{2\lambda}{1-e^{-\lambda t}}\sum_{j=1}^{\infty}\frac{1}{j}(1-e^{-\lambda t})^{j+1} = 2\lambda f(t) = 2\lambda^2 t,$$

we obtain that $g_2(t) = \lambda^2 t^2$, and the second term on the RHS of (10) is thus $(\lambda^2 t^2)/(1-e^{-\lambda t})$.

So, joining all the results, we have that

$$\lim_{t\to\infty}\mathrm{Var}\left(\mathbb{E}\left(\frac{G(t)}{Z(t)}|Z(t)\right)\right) = \lim_{t\to\infty} 4\left(\sum_{k=1}^{\infty}\frac{(1-e^{-\lambda t})^{k-1}}{k^2} + \frac{\lambda^2 t^2}{1-e^{-\lambda t}} - \frac{\lambda^2 t^2}{(1-e^{-\lambda t})^2}\right) = 4\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{2}{3}\pi^2 \quad \square$$

Theorem 1 follows from Eq. (1) using the results in Lemmas 2 and 3.

Acknowledgements

The authors thank Tom S. Weber (WEHI) for contributing to the conjecture of Theorem 1. Part of this work was supported by Science Foundation Ireland grant 12 IP 1263.

References

Biggins, J.D., Grey, D.R., 1997. A note on the growth of random trees. Statist. Probab. Lett. 32 (4), 339–342.
Blitzstein, J.K., Hwang, J., 2014. Introduction to Probability. Chapman and Hall/CRC.
Devroye, L., 1986. A note on the height of binary search trees. J. ACM 33 (3), 489–498.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.
Flajolet, P., Odlyzko, A., 1980. Exploring binary trees and other simple trees. In: 21st FOCS. IEEE, pp. 207–216.
Harris, T.E., 1963. The Theory of Branching Processes. Springer-Verlag, Berlin.
Jagers, P., 1969. Renewal theory and the almost sure convergence of branching processes. Ark. Mat. 7 (6), 495–504.
Kimmel, M., Axelrod, D.E., 2002. Branching Processes in Biology. Springer.
Marchingo, J.M., Prevedello, G., Kan, A.J., Heinzel, S., Hodgkin, P.D., Duffy, K.R., 2016. T cell stimuli independently sum to regulate an inherited clonal division fate. Nature Commun. 7, 13540.
Meli, G., Weber, T.S., Duffy, K.R., Sample path properties of the average generation of a Bellman-Harris process. arXiv:180707031, 2018.
Perié, L., Hodgkin, P.D., Naik, S.H., Schumacher, T.N., de Boer, R.J., Duffy, K.R., 2014. Determining lineage pathways from cellular barcoding experiments. Cell Rep. 6 (4), 617–624.
Pittel, B., 1984. On growing random binary trees. J. Math. Anal. Appl. 103 (2), 461–480.
Pittel, B., 1994. Note on the heights of random recursive trees and random m-ary search trees. Random Struct. Algorithms 5 (2), 337–347.
Resnick, S.I., 2013. Adventures in Stochastic Processes. Springer Science & Business Media.
Robson, J.M., 1979. The height of binary search trees. Aust. Comput. J. 11 (4), 151–153.
Samuels, M.L., 1971. Distribution of the branching-process population among generations. J. Appl. Probab. 8 (04), 655–667.
Weber, T.S., Perié, L., Duffy, K.R., 2016. Inferring average generation via division-linked labeling. J. Math. Biol. 73 (2), 491–523.